On-Line Analytical Mining and Time Series Analysis in Multi-Dimensional Data

A
SYNOPSIS

SUBMITTED IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
IN
COMPUTER SCIENCE

BY

VIJAI DEV

UNDER THE SUPERVISION OF
Prof. Preetvanti Singh
FACULTY OF SCIENCE

DEPARTMENT OF PHYSICS AND COMPUTER SCIENCE
FACULTY OF SCIENCE
DAYALBAGH EDUCATIONAL INSTITUTE
DAYALBAGH, AGRA

April 2017

HEAD
Department of Physics and Computer Science
Faculty of Science

DEAN
Faculty of Science

INTRODUCTION

The rapid pace of change, both in society and in business, can be witnessed all around in technology and information management. This change has resulted in the exponential growth of data that an organization processes and stores. Managers need to understand high volumes of reliable and timely data before they can make the necessary decisions. This data can be accessed through a computerized database. Database is a systematic collection of data that is processed and organized in a way to obtain a relation between the data and can be used jointly by multiple user applications. Many of today's data-enabled applications implement their functionality and business logic in the database itself. The data persisted in a database is often of mission-critical nature and a key asset for the organization as it helps to manage and monitor operational activities, provide decision-support information to bring rapid progress in the organization, and determine a better strategy for the advancement of an organization in the future.

The data warehouse, a special type of database, is one of the most important business intelligence tools for immense data storage, management and processing. Large organizations build data warehouses to *analyze what has occurred within the business across time* in order to obtain a *competitive edge in the marketplace*. Data warehouse is used for analysis of multi-dimensional data. It collects and stores integrated sets of historical data from multiple operational systems and feeds them to one or more data marts. A data warehouse environment includes an extraction, transformation, and loading (ETL) solution, statistical functions, data mining capabilities, and client analysis tools to manage the process of gathering data, transform it into useful and actionable information, and deliver it to business users.

The characteristics of a data warehouse are (Inmon, 2005):

- Data warehouse is designed to analyze data for a particular subject say health care, which makes it subject-oriented
- Data warehouse must put data from disparate sources into a consistent format hence is integrated

- Data warehouse is nonvolatile, i.e. consists of stable data that does not change each time an operational process is executed
- The focus of a data warehouse is to analyze the change over time, i.e. it is time variant

Advantages of Data Warehouse are:

- Data warehouses store credible facts and statistics which a decision maker is able to retrieve as and when needed
- Decision makers can quickly access data from multiple sources in a data warehouse to make quick and accurate decisions
- Data warehouses gather data from different sources and convert it into a single and widely used format which results in accuracy
- Data warehouses stores and integrate historical data and provide facilities regarding advanced query features.
- Users can establish a secure connection to the data warehouse.
- As the stored data has a specific time duration organizations can easily access data for a particular time period
- Data warehouse helps to preserve data for future use

Data Warehouse Architecture

The basic architecture for a data warehouse is as shown in Figure 1. End users directly access data derived from several source systems through the data warehouse.

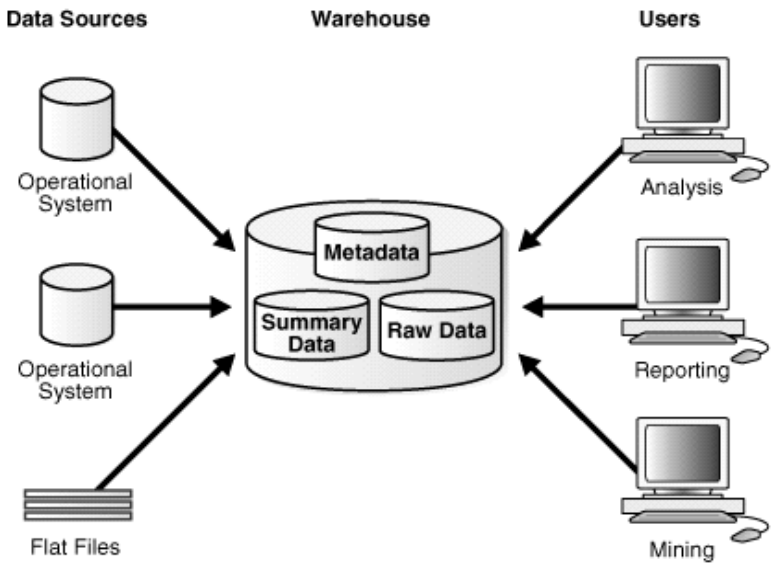

Figure 1: Basic Data Warehouse

The metadata and raw data of a traditional OLTP system is present as an additional type of data, the summary data. Summaries are very valuable in data warehouses because they pre-compute long operations in advance. These can then be used by the decision makers for analysis, reporting or mining.

Data warehouse tools

The information can be extracted from the masses of data stored in a data warehouse by analyzing the data. Popular data warehouse tools for this are:

- Online Analytical Processing (OLAP)
- Data Mining
- A wide array of statistical functions, including descriptive statistics, hypothesis testing, correlation analysis, test for distribution fit, cross tabs with Chi-square statistics, and ANOVA.

1.1 ONLINE ANALYTICAL PROCESSING (OLAP)

OLAP is a technology used to organize large business databases and support business intelligence. OLAP databases are divided into one or more cubes, and each cube is organized and designed by a cube administrator to fit the way that data can be easily retrieved and analyzed.

A cube is the data structure that aggregates the measures by the levels and hierarchies of each of the dimensions that can be analyzed. For example, a cube combines time, geography, and product lines with summarized data, such as sales or inventory figures.

The source data for OLAP is Online Transactional Processing (OLTP) databases that are commonly stored in data warehouses. OLAP data is derived from this historical data and is aggregated into structures that are useful in sophisticated analysis.

OLAP databases contain two basic types of data:

- **Measures**: are a set of values in a cube that are based on a column in the cube's fact table and are usually numeric values. Measures are the central values in the cube that are preprocessed, aggregated, and analyzed;
- **Dimensions**: are a set of one or more organized hierarchies of levels in a cube that a user understands and uses as the base for data analysis.

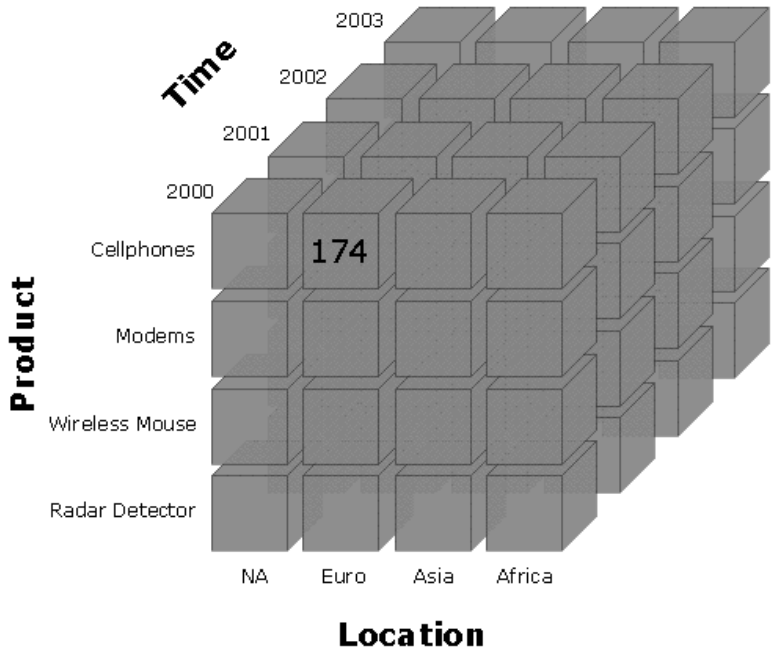

Figure 2: Example of OLAP cube (*Source: http://www.hypertextbookshop.com/dataminingbook*)

The figure shows a data cube with a measure and 3 dimensions: time, location and product.

Different types of OLAP are:

- **Relational OLAP (ROLAP)**: ROLAP systems work primarily from the data that resides in a relational database, where the base data and dimension tables are stored as relational tables. This model permits multidimensional analysis of data.
- **Multidimensional OLAP (MOLAP):** MOLAP is commonly considered as the classic form of OLAP. One of the major distinctions of MOLAP against a ROLAP is that data are pre-summarized and are stored in an optimized format in a multidimensional cube, instead of in a relational database. In MOLAP model, data are structured into proprietary formats in accordance with a client's reporting requirements with the calculations pre-generated on the cubes.
- **Hybrid OLAP (HOLAP):** HOLAP incorporates the best features of MOLAP and ROLAP into a single architecture. HOLAP systems store larger quantities of detailed data in the relational tables while the aggregations are stored in the pre-calculated cubes. HOLAP also has the capacity to *drill through* from the cube down to the relational tables for delineated data.
- **Web-Enabled OLAP (WOLAP):** WOLAP pertains to OLAP application which is accessible via the web browser. Unlike traditional client/server OLAP applications, WOLAP is considered to have a three-tiered architecture which consists of three components: a client, a middleware and a database server.
- **Desktop OLAP (DOLAP):** DOLAP permits a user to download a section of the data from the database or source, and work with that dataset locally, or on their desktop.
- **Mobile OLAP (MOLAP):** Mobile OLAP enables users to access and work on OLAP data and applications remotely thorough the use of their mobile devices.
- **Spatial OLAP (SOLAP):** SOLAP includes the capabilities of both Geographic Information Systems (GIS) and OLAP into a single user interface. It facilitates management of both spatial and non-spatial data.

1.2 DATA MINING

OLAP provides a view of what is happening, but cannot predict what will happen in the future or why it is happening. Data Mining, on the other hand, is a combination of discovery and prediction techniques. Data mining uses large quantities of data to create models which provides insights that are revealing, significant, and valuable. Data can be used, for example, to:

- Predict the customers likely to change service providers
- Discover the factors involved with a disease
- Identify fraudulent behavior.

Data Mining Techniques include:

- Classification: Grouping items into discrete classes and predicting which class an item belongs to; the algorithms are Decision Tree, Naive Bayes, Generalized Linear Models (Binary Logistic Regression), and Support Vector Machines.
- Regression: Approximating and predicting continuous numeric values; the algorithms are Support Vector Machines and Generalized Linear Models (Multivariate Linear Regression).
- Anomaly Detection: Detecting anomalous cases, such as fraud and intrusions; the algorithm is one-class Support Vector Machines.
- Attribute Importance: Identifying the attributes that have the strongest relationships with the target attribute; the algorithm is Minimum Descriptor Length.
- Clustering: Finding natural groupings in the data that are often used for identifying customer segments; the algorithms are k-Means and O-Cluster.
- Associations: Analyzing items that are likely to be associated; the algorithm is priori.
- Feature Extraction: Creating new attributes (features) as a combination of the original attributes; the algorithm is Non-Negative Matrix Factorization.

1.3 TIME SERIES ANALYSIS

In a time-series analysis, time is the main variable used to describe the data under analysis. Time series analysis is associated with the discovery and use of patterns (such as periodicity, seasonality, or cycles), and prediction of future values. Time series forecasting is an important area of forecasting in which past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship. The model is then used to extrapolate the time series into the future. This modeling approach is particularly useful when little knowledge is available on the underlying data generating process or when there is no satisfactory explanatory model that relates the prediction variable to other explanatory variables.

Time-series patterns can be described in terms of two basic concepts:

- Trend: This concept represents a general systematic linear or nonlinear component that changes over time. It does not repeat, or at least it does not repeat within the time range captured in the data.
- Seasonality: This can be identified by regularly spaced peaks and troughs. These have a consistent direction and approximately the same magnitude every year which is relative to the trend.

1.4 ORACLE OLAP

Oracle OLAP is a world class multidimensional analytic engine embedded in Oracle Database. Oracle OLAP cubes deliver sophisticated calculations using simple SQL queries which produce results with speed of thought response times. Oracle OLAP allows centralized management of data and business rules in a secure, scalable and enterprise-ready platform. It makes it easy to produce analytic measures, data mining and time-series calculations, financial models, forecasts, allocations, and regressions.

Features of OLAP include: Integration of multidimensional technology; Ease of administration; Tools for creating and managing dimensional objects; Querying dimensional objects; and Efficient storage and uniform availability of summary data

LITERATURE REVIEW

In today's challenging times good decision-making becomes critical as this requires taking into consideration all the relevant data available. The best possible source for such data is a well-designed data warehouse. The concepts of data warehouse can be studied in Watson (2002), Sahama & Croll (2007), Ifeanyi et al. (2014), Namrata & Singh (2014) and Hu, Li, Hu & Yang. (2015). Data warehouse has emerged as an important area of study for both industrial and research community reflecting the magnitude and impact of data-related problems to be solved in business organizations.

Pedersen & Jensen (1998, 1999) presented many new research challenges posed by clinical data warehouse to be met by the database research community. Berndt, Hevner, & Studnicki (2003) focused on the technical challenges of designing and implementing an effective data warehouse for health care information. Illustrations of actual data designs and reporting formats from the CATCH data warehouse were used throughout the discussion. Ramamurthy, Sen, & Sinha (2008) proposed and empirically tested a comprehensive model as a key determinant of data warehouse adoption. Salguero, Araque, & Delgado (2008) described a spatio-temporal extension of an ontology language which facilitates the generation of the scheme of the data warehouse as well as the design of the processes which extract, transform and load the data from the sources in the data warehouse according to the temporal characteristics of the data sources. Zhou et al. (2008) introduced a data warehouse system, which is based on the structured electronic medical record system and daily clinical data for Traditional Chinese Medical clinical researches and medical knowledge discovery. The system consists of several key components: clinical data schema, ETL tool, OLAP based on business objects, and integrated data mining functionalities.

Danubianu, Socaciu, & Barila (2009) studied the possibility and necessity to deploy a data warehouse for the tourism industry that aims to support decision makers by giving them different views for the same piece of data. Gonzalez et al. (2010) modeled the RFID data warehouse using a movement graph-centric

view, which makes the warehouse conceptually clear, better organized, and helps in obtaining significantly deeper compression and performance gain over competing models in the processing of path queries. Zhou et al. (2010) introduced a clinical data warehouse system, which incorporates the structured electronic medical record (SEMR) data for medical knowledge discovery and clinical decision support. High amount of relevant information is contained in reports stored in the electronic patient records and associated metadata. Cuggia et al. (2011) developed a medical search engine R-oogle for clinicians. The system consists of a data warehouse (full-text reports and structured data) imported from two different hospital information systems.

Boussadi et al. (2012) evaluated the use of a clinical data warehouse coupled with a clinical information system to test and refine alerts for medication orders control. A clinical decision rule refinement process was used to assess alerts. Evans, Lloyd & Pierce (2012) presented a framework to create a number of decision support applications that are dependent on the enterprise-wide data warehouse in addition to a patient's current information in the electronic medical records. Mul et al. (2012) reported in-house development of an integral part of the data warehouse specifically for the intensive care units. It was modeled using Atos Origin Metadata Frame method. Burstein, Silva, Jelinek & Stranieri (2013) described the components of an architecture, which includes a robust data warehouse as an infrastructure for comprehensive clinical knowledge management. The proposed DD-DGMS architecture incorporates the dynamic dimensional data model as its elemental core. Roelofs et al. (2013) combined a Computer Aided Theragnostics data warehouse with automated tools for feature extraction. Hamad & Qader (2014) integrated data warehouse and knowledge warehouse to develop a knowledge-driven DSS for managing (i.e. storing and retrieving) the knowledge to improve the process of decision making and management of the market resources. Krasowski et al. (2014) developed a data warehouse collaboratively between an academic medical center and a private company. The data warehouse contains data from the EHR, LIS, admission/discharge/transfer system, and billing records which can be accessed using a self-service data access tool Starmaker. Ross et al. (2014) highlighted the HMORN VDW data model, its governance

principles, data content, and quality assurance procedures. The goal was to share the data model and its operations to those wishing to implement a distributed interoperable health care data system. Kamble, Desai & Vartak (2015) developed a data warehouse for supply chain. An improved sales forecasting model was presented by the authors based on kernel based support vector machine regression.

Object-Oriented Data warehouse

Object-oriented modeling has emerged as an important technique for database design as it organizes system as collection of interacting objects that combine data and behavior (Gandhi and Jain, 2011).

Ravat & Teste (2000) dealt with object-oriented data warehouse design by integrating temporal and archive data. Functions were provided allowing the administrator to specify a data warehouse from a global source schema. Abelló & Martin (2003) developed a data warehouse for object-oriented temporal data. Cheng, Yang, and Lin (2004) adopted distributed object and mobile object technologies, RosettaNet implementation framework, as well as Holon and holarchy concepts derived from studying social organizations and living organisms. The generic Holon was first developed by adopting the technologies of the distributed object-oriented approach with common object request broker architecture infrastructure, client/server architecture, a knowledge base, and data warehousing to achieve the properties of Holon. Peng, Li, Feng, Li & Liu (2005) developed a framework using Smalltalk to prepare data for data warehousing, in which an object deputy model and database connecting tools have been implemented to provide an easy-to-use way to resolve inconsistency and conflicts. Ali, Soh, & Torabi (2006) implemented an agent-oriented framework to make the process of incorporating business rules into business process less complex. Sarkar, Choudhury, Chaki, & Bhattacharya (2009) proposed a Graph semantic based Object Oriented Multidimensional Data Model (GOOMD) which defines a set of graph based formal constructs to specify the conceptual level design of data warehouses. Alaskar & Shaikh (2009) introduced the UML Profile for Modeling DWH on a conceptual level. It uses features of UML

intended for the purpose of creating abstract, general models. The UML Profile is applied to illustrate Hajj pilgrims' private tour.

Gosain & Mann (2010) defined an object oriented multi-dimensional data model which includes aggregation, generalization, multiple path hierarchies and multiplicity. Different operators necessary to make a query and format results were also described. Singh and Singh (2010) identified the reasons for data deficiencies, non-availability or reach ability problems at all the stages of data warehousing and formulated descriptive classification of these causes. Soni, Ansari, Sharma & Soni (2011) surveyed of current techniques of knowledge discovery in databases using data mining techniques that are in use in medical research particularly in Heart Disease Prediction. Gosain & Mann (2014) presented the empirical validation of the metrics for multidimensional data warehouse models at conceptual level. Quality attributes, understandability and efficiency were evaluated through various combinations of metrics.

Object –relational data warehouse

Multimedia data are semi-structured or non-structured including videos, audios, graphics and text which are best served by object-relational technologies. In addition to its traditional role in the safe and efficient management of relational data, object-relational database also provides support for complex objects.

Eder, Frank, Morzy, Wrembel & Zakrzewicz (2000) presented a research project aiming at the design and development of an Object–Relational Data Warehousing System (ORDAWA). Cooley et al. (2008) developed Themodel repository (MREP), was a relational database management system (RDBMS), under the auspices of models of infectious disease agent study (MIDAS). The model repository is comprised of: a source code repository and version control system; a model documentation tree; a data warehouse; and an API consisting of a database MREP. Liu & Lai (2008) proposed a conceptual model for designing an object-relational data warehouse, and provided a semi-automated methodology for deriving the model from the standard documentation of the object-relational database. Pahwa & Chhabra (2014) described a

technique for designing a relational schema from an object model and then transforming it into data warehouse. Loyola, Sepulveda & Hernandez (2015) developed methods to improve response time for queries associated with a data warehouse design for Changing Dimensions tables, and compared response times for SCD data in Relational Database and Object-Relational. The research focuses on proposing a general design for Object-Relational of Slowly Changing Dimensions (SCD) data types. Chhabra, Kumar & Pahwa (2016) illustrated a technique for designing a relational schema from an object model, representing in their UML form and then transforming it into data warehouse.

Query optimization

Query optimization is the bottleneck of database application performance especially those which store history i.e. data warehouse. In order to reduce the complexity of the query generation process and in order to preserve portability to other database systems query optimization architecture is very useful. A query optimizer determines the best strategy for performing each query. The concepts of Query Optimization are discussed by Johnson & Srivatsa (2012), Arpitha (2013) and Patel & Patel (2015).

Karde & Thakare (2010) proposed a tree based materialized view selection algorithm for query processing. Authors also proposed node selection algorithm for fast materialized view selection in distributed environment. Sukheja & Singh (2010) focused on query optimization technique which generates sequences of SQL statements in order to produce the requested information. Semantic query optimizer architecture was suggested for these applications. Bajda-Pawlikowski, Abadi, Silberschatz & Paulson (2011) discussed performance-oriented query execution strategies for data warehouse queries in split execution environments, with particular focus on join and aggregation operations. Fender & Moerkotte (2012) presented an algorithm for cyclic query graphs and compared it with the best top-down and bottom-up algorithms.

Alexandrov et al. (2014) presented Stratosphere, an open-source software stack for parallel data analysis. Stratosphere brings together a unique set of features that allow the expressive, easy, and efficient programming of analytical applications at very large scale. Badia & Wagner (2014) proposed a logical framework to analyze complex predicates (those involving a subquery) in SQL. A new operator was proposed in the relational algebra for handling such predicates, and its properties were studied. Li, He, Yan, & Safiullah (2014) proposed to augment SQL with set predicate to bring out otherwise obscured set-level semantics. Two approaches were presented for processing set predicates—an aggregate function-based approach and a bitmap index-based approach. A histogram-based probabilistic method of set predicate selectivity estimation was designed for optimizing queries with multiple predicates. Qiao, Cheng, Chang, & Yu (2014) analyzed the factors that affect the accuracy of distance estimation in landmark embedding. Two optimization techniques on graph compression and graph online search were also proposed with the goal of further reducing index size and improving query accuracy. Rahul & Janardan (2014) proposed a technique to solve any top-k GIQ problem efficiently. The technique relies only on the availability of an efficient solution for the underlying (non-top-k) GIQ problem. Diallo, Rodrigues, Sene & Lloret (2015) surveyed the state of the art of the techniques used to manage data and queries in wireless sensor networks based on the distributed paradigm. A classification of these techniques is also proposed. Fan, Zhang, Kok, Lu & Ooi (2015) studied the query optimization problem in declarative crowdsourcing systems which were designed to relieve the user from the burden of dealing with the crowd. Khachatryan, Müller, Stier & Böhm (2015) demonstrated how to improve a self-tuning method significantly by starting with a carefully chosen initial configuration and proposed initialization by dense subspace clusters in projections of the data, which improves both accuracy and robustness of self-tuning. Myalapalli & Dussa (2015) proposed optimizing techniques for ensuring high performance data warehousing which could serve as a tuning/bench-marking/management tool for overhauling data warehouse querying practices and processing. Ramachandra, Chavan, Guravannavar & Sudarshan (2015) addressed the issue of automatically transforming a program written assuming synchronous query submission to one that exploits asynchronous query submission. Xu, Tu, & Wang (2015) built a series of

physical models for energy estimation of individual relational operators based on their resource consumption patterns. As the execution of a query plan is a combination of multiple relational operators, authors used the physical models as a basis for a comprehensive energy model. Brodsky, Shao & Riddick (2016) introduced National Institute of Standards and Technology (NIST)'s Sustainable Process Analytics Formalism (SPAF) to facilitate the use of simulation and optimization technologies for decision support in sustainable manufacturing. SPAF allows formal modeling of modular, extensible, and reusable process components and enables sustainability performance prediction, what-if analysis, and decision optimization based on mathematical programming.

OLAP

On-line analytical processing (OLAP) describes an approach to decision support, which aims to extract knowledge from a data warehouse, or more specifically, from data marts. Its main idea is providing navigation through data to non-expert users, so that they are able to interactively generate ad hoc queries without the intervention of IT professionals (Abelló et al., 2015).

Nguyen, Schiefer & Tjoa (2005) introduced an enhanced business intelligence architecture using OLAP that covers the complete process to sense, interpret, predict, automate and respond to business environments and thereby aims to decrease the reaction time needed for business decisions. Stolba & Tjoa (2006) proposed a data warehouse based approach as a suitable solution for the integration of external evidence-based data sources into the existing clinical information system and data mining techniques for finding appropriate therapy for a given patient and a given disease. Through integration of data warehousing, OLAP and data mining techniques in the healthcare area, an easy to use decision support platform, which supports decision making process of care givers and clinical managers, is built. Sahay & Ranjan (2008) identified the need for real time business intelligence (BI) in supply chain analytics. The authors focused on the necessity to revisit the traditional BI concept that integrates and consolidates

information in an organization to support firms that are service oriented and seeking customer loyalty and retention.

Kehlenbeck & Breitner (2009) presented an innovative approach based on standard Semantic Web technologies. This approach facilitates the exchange of business calculation definitions and allows for their automatic linking to specific data warehouses through semantic reasoning. A novel standard proxy server which enables the immediate application of exchanged definitions was also introduced. Ordonez & Chen (2009) computed parametric statistical tests treating patient records as elements in a multidimensional cube. Authors introduced a technique that combines dimension lattice traversal and statistical tests to discover significant differences in the degree of disease within pairs of patient groups.

Chaki & Sarkar (2010) proposed an analytical model using Petri Net for distributed data management in a data warehouse to ease the OLAP operations. Some of the properties of the model like safeness, boundedness, liveness and conservativeness are also verified. Cuzzocrea & Gunopulos (2010) proposed a novel framework for efficiently computing and querying multidimensional OLAP data cubes over probabilistic data, which well-capture previous kinds of data. Several models and algorithms supported in the proposed framework were presented and described based on well-understood theoretical statistical/probabilistic tools. Grabova, Darmont, Chauchat, & Zolotaryova (2010) reviewed web-based business intelligence approaches. Authors also discussed the existing approaches and tools working in main memory and/or with web interfaces (including freeware tools), relevant for small and middle-sized enterprises in decision making. Cuzzocrea (2011) introduced a novel framework for estimating OLAP queries over uncertain and imprecise multidimensional data streams, along with three relevant research contributions: developing a probabilistic data stream model, developing a possible-world semantics for uncertain and imprecise multidimensional data streams, and developing an innovative approach for providing theoretically-founded estimates to OLAP queries over uncertain and imprecise multidimensional data streams that exploits the well-recognized probabilistic estimators theory. Li &

Zhou (2011) studied some key issues of smart grid, such as distributed cooperation and control, data and application integration, and knowledge-based comprehensive decision Authors also discussed the use of data warehouse and OLAP for smart grid. Park, Yu, Park & Kim (2013) studied a NetCube, which is a comprehensive network traffic analysis model based on multidimensional OLAP data cube

Data Mining

Data mining involves determining pattern from or fitting models to observed data. It is sophisticated data analytical method that focuses upon exploration and devoloping new insights for supporting decision making. This extracted information is useful for identifying trends, forming a prediction or classification model and summarizing a database.

Palaniappan & Lin (2008) presented a prototype clinical decision support system which combines the strengths of both OLAP and data mining. It provides a rich knowledge environment which is not achievable by using OLAP or data mining alone. Srinivas, Rani & Govrdhan (2010) examined the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. Peng, Zhang, Tang & Li (2011) proposed an incident information management framework that consists of three major components: a high-level data integration module in which heterogeneous data sources are integrated and presented in a uniform format; a data mining module to identify useful patterns for pre-incident and post-incident information management; and a multi-criteria decision-making module to assess the current situation, find the satisfactory solutions, and take appropriate responses in a timely manner.

Sharma & Mehta (2012) discussed the characteristic computational needs of agriculture data which is essentially seasonal and uncertain along with some suggestion regarding the use of data mining techniques as a tool for knowledge management in agriculture. Panov, Soldatova & Dzeroski (2014) presented Onto-core that defines the most essential data mining entities in a three-layered ontological

structure comprising of a specification, an implementation and an application layer. Sim, Choi & Kim (2014) developed a data mining approach to which large amounts of trace data are inputted to infer fault-introducing machines in the form of a L⇒R rule, where R contains the fault type and L contains a machine sequence that is the primary cause of the fault type.

Kumar & Toshniwal (2015) proposed a framework that used K-modes clustering technique as a preliminary task for segmentation of a large number of road accidents on road network of Dehradun (India). Association rule mining was used to identify the various circumstances that are associated with the occurrence of an accident for the entire data set and the clusters identified. Ram & Doegar (2015) used data mining techniques for the classification of Statlog heart disease datasets. These supervise machine learning algorithms were compared on the basis of classification accuracy and performance matrices.

Castro & Kim (2016) used data mining classification models to detect factors with the greatest influence on car accidents. The experimental objective explored the role of different factors on injury risk using a Bayesian network, decision tree and ANN. Dimitriado, Papaemmanouil & Diao (2016) developed an Automatic Interactive Data Exploration (AIDE) framework that assists users in discovering new interesting data patterns and eliminating expensive ad-hoc exploratory queries. Dindarloo & Siami-Irdemoosa (2017) explored the application of classification and clustering approaches for pattern recognition and failure forecasting on mining shovels. The shovels were classified into four clusters using k-means clustering algorithms. Kumar & Toshniwal (2016) applied k-means algorithm to group the accident locations into three categories, high-frequency, moderate-frequency and low-frequency accident locations. Mirge, Verma & Gupta (2016) presented a novel technique for excavating heavy traffic flow patterns in bi-directional road network, Tang, He, Baggenstoss & Kay (2016) introduced a Bayesian classification approach for automatic text categorization by using class-specific features. Unlike conventional text categorization approaches, authors proposed a method to select a specific feature subset

for each class. Taylor et al. (2016) presented a data mining methodology for driving condition monitoring via controller area network-bus data based on the general data mining process. The approach is applicable to many driving condition problems .The example of road type classification without the use of location information was investigated. Zhang, Zhang, Liu & Liu (2016) introduced a multi-task multi-view clustering framework which integrates within-view-task clustering, multi-view relationship learning, and multi-task relationship learning.

Time series

The usage of time series models is in obtaining an understanding of the underlying forces and structure that produce the observed data, and fit a model and proceed to forecasting, monitoring or even feedback and feed-forward control. Time series occur frequently for industrial data. Yang and Zou (2001) described use of Statistical models in time series data analysis and forecasting.

Liu, Bhattacharyya, Sclove, Chen & Lattyak (2001) used a fast-food restaurant franchise as a case to illustrate how data mining can be applied to a time series. Time series data mining at both, the store level and corporate level were, discussed. Lim & McAleer (2002) used Box-Jenkins ARIMA model to forecast tourist arrival to Australia from Hong Kong, Malaysia and Singapore. Goh & Law (2002) applied SARIMA and MARIMA time series models with interventions in forecasting tourism demand using ten arrival series for Hong Kong. Yorucu (2003) evaluated forecast accuracy by using tourism arrivals data to North Cyprus and Malta using the forecasting methods: Actual Static (AS) (econometric), Double Exponential Smoothing (DES), Holt Winters (HW) and Autoregressive Moving Averages (ARMA). Shitan and Wee (2003a, b) updated and compared the performance of three time series models for modeling tourist arrivals to Malaysia. One of them is within the class of ARMA models and the other two are in the class of ARFIMA models.

Chu (2004) examined the accuracy of a forecasting model in predicting international tourism arrivals. The cubic polynomial model was employed to forecast the volume of tourist arrivals. Naudé & Saayman (2005) explained the uses of both, cross-section data as well as panel data, to identify the determinants of tourism arrivals in 43 African countries, taking into account the country of origin of tourists. Patelis, Petropoulos, Nikolopoulos, Lin & Assimakopoulos (2005) proposed an e-government decision support system for tourism demand analysis and forecasting. Papatheodorou and Song (2005) presented diagrammatic and forecasting analysis of tourism both world-wide and at regional level. Chen, Bloomfield & Cubbage (2007) described the use of three major U.S. national parks as applications of statistically selecting appropriate methods to forecast attendance.

Chu (2008, 2009) defined the ARAR model and its usefulness as a forecast-generating mechanism for tourism demand for nine major tourist destinations in the Asian-Pacific region. Fernandes and Teixeira (2008) developed time series models, and applied these to sensitivity studies in order to predict the tourism demand. Hilaly et.al (2008) developed forecasting techniques for tourism demand prediction. The study provides different econometric models used in modeling and forecasting tourism demand. Song and Li (2008) integrated both qualitative and quantitative forecasting approaches for improving forecasting accuracy. Viswanathan, Widiarta & Piplani (2008) investigated the effectiveness of top-down and bottom-up forecasting strategies for estimating an aggregate time series when the underlying pattern of the sub-aggregate time series is intermittent. Chaiboonsri & Chaitip (2009) focused on forecasting methods based on X-12-ARIMA seasonal adjustment and produced forecasts of international tourist arrivals to India. Ismail, Suhartono, Yahaya & Efendi (2009) described the time series data with extreme change in its mean caused by an intervention which comes from external and/or internal factors. Witt (2009) discussed seasonal patterns of tourism demand and concluded that the effects of various influencing factors on this demand tend to change over time.

Athanasopoulos, Hyndman, Song & Wu (2010) evaluated the performance of various methods for forecasting tourism data. The data included 366 monthly series, 427 quarterly series and 518 annual series, all supplied by tourism bodies or by academics that had used them in previous tourism forecasting studies. Chaovanapoonphol, Lim, McAleer & Wiboonpongse (2010) investigated the relationship between the demand for international tourism to Thailand and its major determinants. The time series of tourist arrivals and economic determinants were examined using ARIMA with exogenous variables (ARMAX) models to analyze the relationships between tourist arrivals from different countries to Thailand.

Chujai, Kerdprasop & Kerdprasop (2013) developed a model to forecast the electricity consumption in a household and to determine the most suitable forecasting period (daily, weekly, monthly, or quarterly). Ikponmwosa & Aliu (2014) studied students' patronage of the e-library (E-Learning Centre) in John Harris Library, University of Benin as a function of time and made predictions and recommendations for future occurrences using time series models. Kumar & Anand (2014) used a time series modeling approach (Box-Jenkins' ARIMA model) to forecast sugarcane production in India. The order of the best ARIMA model was found to be (2,1,0). Jebb, Tay, Wang & Huang (2015) introduced time series analysis to psychological research, an analytic domain that has been essential for understanding and predicting the behavior of variables across many diverse fields. Different time series modeling techniques were surveyed to address various topics of interest to psychological researchers.

Atto, Trouvé, Nicolas, & Lê (2016) described the statistical properties of wavelet operators when the observation model can be seen as the product of a deterministic piecewise regular function (signal) and a stationary random field (noise). This multiplicative observation model is analyzed using time-series analysis. Bose, Kasabov, Bruzzone, & Hartono (2016) presented spiking neural networks (SNNs) for remote sensing spatiotemporal analysis of image time series, which make use of the highly parallel and low-power-consuming neuromorphic hardware platforms possible. Authors illustrated this concept with

the introduction of the first SNN computational model for crop yield estimation from normalized difference vegetation index image time series. Chen et al. (2016) presented a novel space affine matching feature by introducing the time domain and frequency domain features. The time domain feature is used to discern different stimuli, while the frequency domain feature is used to eliminate the delay. Cherrington et al. (2016) compared VI data for tropical evergreen forests in three zones north of the equator (the Guianas, central Africa, and northern Borneo) using time-series analysis. Kolozali, Puschmann, Bermudez-Edo & Barnaghi (2016) described a framework for real-time semantic annotation and aggregation of data streams to support dynamic integration into the Web using the advanced message queuing protocol. Ye, Mistry, Bouguettaya & Dong (2016) proposed a cloud service composition framework that selects the optimal composition based on an end user's long-term Quality of Service (QoS) requirements. The proposed framework uses a new multivariate QoS analysis to predict the long-term QoS provisions from service providers' historical QoS data and short-term advertisements represented using time series analysis.

Fattahi, Agram & Simons (2017) developed a workflow to estimate the time series of azimuth mis-registration using a network-based enhanced spectral diversity (NESD) approach in order to reduce the impact of temporal decorrelation on coregistration. Vans, Staals, Loffler, Dykes & Speckmann (2017) proposed a geometric model for trend-detection in one-dimensional time-varying data, inspired by topological grouping structures for moving objects in two- or higher-dimensional space. Shnitzer, Talmon & Slotine (2017) proposed a method for building an intrinsic representation of signals in a purely data-driven manner by applying manifold learning technique, diffusion maps, to learn the intrinsic model of the latent variables of the dynamical system solely from the measurements. Yokoya, Zhu & Plaza (2017) presented a new framework, multisensory coupled spectral unmixing (MuCSUn) that solves unmixing problems involving a set of multisensor time-series spectral images in order to understand dynamic changes of the surface at a subpixel scale.

Motivation of proposed work

From the above discussion following are the main observations:

- In today's business world a plethora of data exists in repositories distributed across the globe, crossing institutional, regional and national boundaries. To be able to harness this data and move it across these boundaries has the potential to provide great scientific and business insight, to the benefit of many protagonists in industrial community (Stell, Sinnott &Ajayi, 2006).

  Developing a Data Warehouse facilitates efficient storage, enhances timely analysis and increases the quality of real time decision making processes. Data warehouse provides information to users in areas ranging from research to management. Data warehouse has great importance from the real-life application viewpoint in advancement of the service provided and in increasing the business opportunities.

- The traditional data warehouse provides only numeric and character data analysis. But as information technologies progress, complex data such as semi-structured and unstructured data become vastly used (Sahama and Croll, 2007)

  Development of new data warehouse technologies is required to provide complex data analysis, knowledge discovery and decision making support.

- Complex data types including semi-structured or non-structured such as video, audio, graphic and text are best served by object-relational technologies (Stonebraker & Brown, 1999)

In addition to its traditional role in the safe and efficient management of relational data, an object-relational database provides support for complex objects. In addition to multimedia data, complex objects also provide multiple base atomic types and user defined object types such as composite attributes, collection sets or variable arrays.

- Today time series data are being generated at an unprecedented speed from almost every application domain, for example daily fluctuations of stock market, traces of dynamic processes and scientific experiments, medical and biological experimental observations, various readings obtained from sensor networks, position updates of moving objects in location-based services etc. (Wang et al., 2013).

  Time series data occurs frequently in business applications and in science. There has been a lot of interest in querying and mining such data which results in development of new methodologies for indexing, classification, clustering and approximation of time series.

- **It was also observed that data mining techniques have been applied widely in business world, but are relatively new to higher education (Chang, 2006). Education is an essential element for progress of a country (Bhise, Thorate & Sukekar, 2013).**

  Education data mining is an emerging research field that allows data mining in education environment. It can be used to analyze student behavior and student performance, and also to enhance teaching and learning.

PROPOSED WORK

The focus of this research study is to develop OLAP mining and time series analysis tools for multi-dimensional data (data warehouse) to improve the knowledge discovery for providing much useful and powerful decision.

Objectives of the study

1. Incorporating an object-relational data warehouse in decision support system
2. Development of a OLAP tool for time series analysis
3. Querying time series data
4. Query optimization

Variables

This study will deal with the variables based on the application area chosen for the study area (Education System, Health Care or Tourism). Two types of attributes: categorical attribute corresponding to nominal, binary and ordinal variables; and continuous attribute corresponding to integer, interval-scaled and ratio-scaled variables will be considered.

Methodology

There are two approaches to develop a data warehouse:

- *Top-down approach: The top-down approach* is used when the technology and the economic problems are well known. This approach achieves the synergy between the business subjects. It is a systemic method which minimizes the integration problems, but is expensive, and has a low flexibility.

- *Bottom-up approach: The bottom-up approach* is based on experiments and prototypes. It is a flexible method that allows the organization to go further with lower costs, to build independent data marts and to evaluate the advantages of the new system as they go along.

The two approaches may be combined to benefit by the advantages provided by each of them.

From software engineering point of view, one of the following models will be used:

- *V-Model*, which requires a structured and systematic analysis at each step, before going forward;
- *Spiral Model*, which allows fast generation of more and more developed functional systems. It also allows an efficient management of the users' requirements and reduces the possible risks.

Stages of building proposed system

1. **Development of a feasibility study**: This stage starts with a strategic analysis, including the evaluation of organization business lines. The data warehouse will be built in a number of designing, developing and refining iterations according to the tactical and strategic business requirements.

2. **Business lines analysis:** This is an important stage in the data warehouse development cycle. Its main purpose is business understanding and business requirement identification. The users of a data warehouse form a heterogeneous group and they will formulate varied requirements, which only partly can be foreseen in data warehouse development stage.

On the basis of requirements, the business subject priorities will be established, depending on their relative importance, the costs and feasibility of the necessary data. This priority list will used to establish the subsequent iterations of data warehouse development.

3. **Data warehouse architecture design**: First, the data warehouse *logical architecture* will be defined. The required data may be collected as a central repository storing the data of the entire organization; an optional operational data store, or as one or more data marts and one or more metadata repositories.

   The *data architecture* will organize the data sources and collections and will define the quality and management standards, both for data and metadata.

   The *application architecture* will present the software components to provide the implementation of the business functionality within the data warehouse, as well as the data transfer from its source to users, that is data extracting, cleaning, transforming, loading, refreshing and accessing.

   The *technical architecture* will provide the infrastructure for data and application architectures. It will include the server, hardware and software components for connection and communication, and the users' workstations.

   The *support architecture* will include tools for backup/recovery, archiving, performance monitoring, as well as the organizational functions necessary for the technological investment management.

4. **Selection of the technological solution:** In this stage the possible tools for implementing data and application architecture will be identified to provide technical and support architecture functions.

5. **Planning the project iterations:** The data warehouse is implemented one subject area at a time, depending on the priorities established in the stage of business requirement identification. In this stage, the identified business and technical requirements will be refined for data warehouse development and implementation.

   The preliminary analysis made in the previous stages will be detailed. All the constraints imposed by the source systems will be identified.

6. **Data warehouse modeling:** In this stage the data warehouse physical model (database schema) will be achieved, metadata will be defined and data source list will be updated to include all the information necessary for the implementation.
7. **Data warehouse testing and implementation:** Once the planning and design stages are completed, the current iteration for data warehouse implementation will start. In this stage, the development and testing environments will be established, the hardware and software components will be installed and the configuration management process will be implemented.

The developed system will be implemented for admission process of an institute in general, and Dayalbagh Educational Institute (Deemed University) in particular as a case study.

The architecture of the proposed system will be as shown in Figure 3.

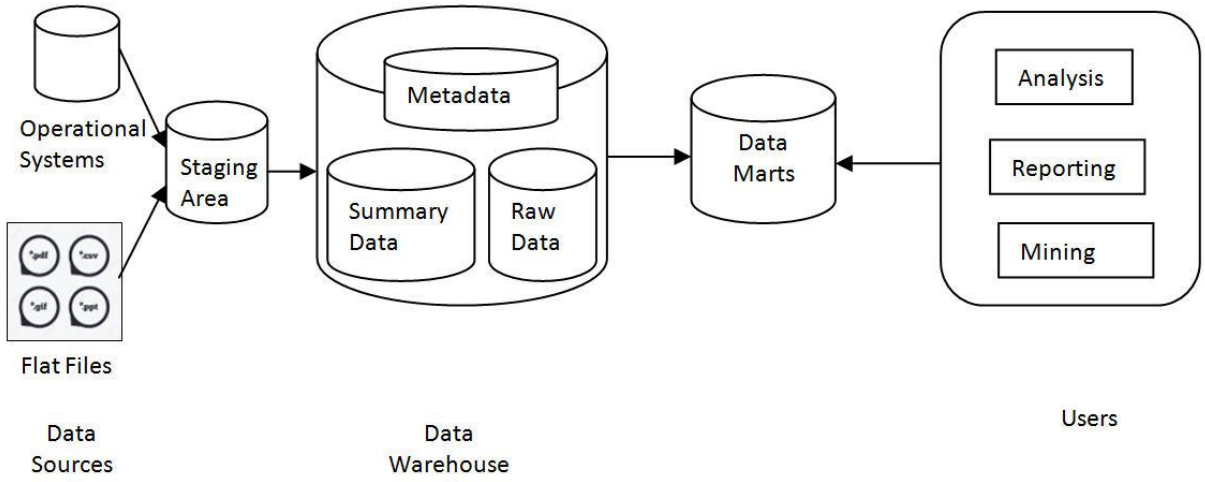

Figure 3: The Architecture of a data Warehouse

The functional components are shown in Figure 4 and the information pipeline for the data integration and processing is shown in Figure 5.

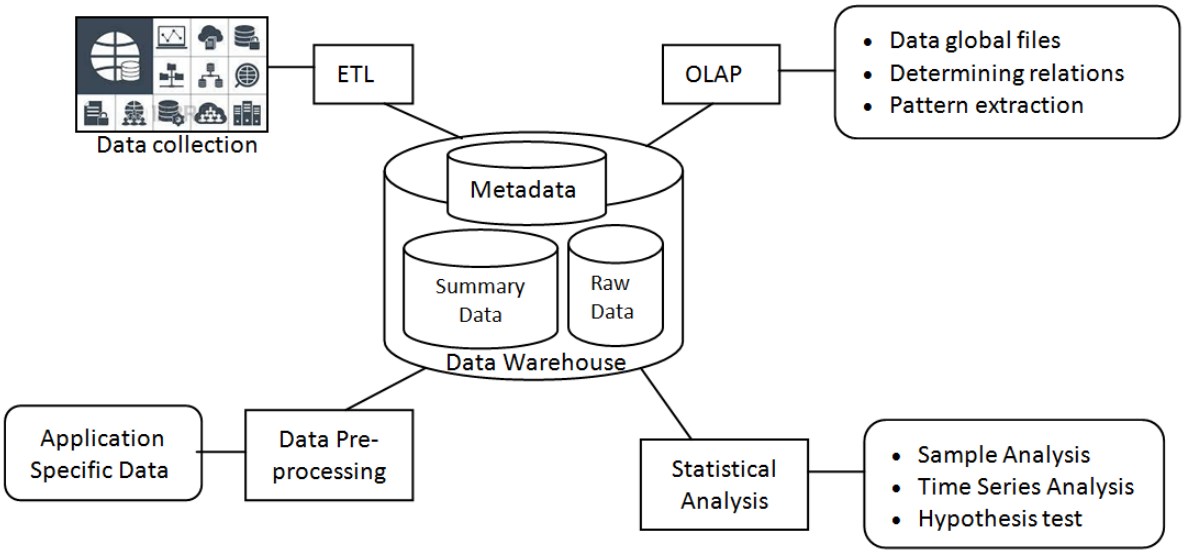

Figure 4: The Functional Component of Proposed Data Warehouse

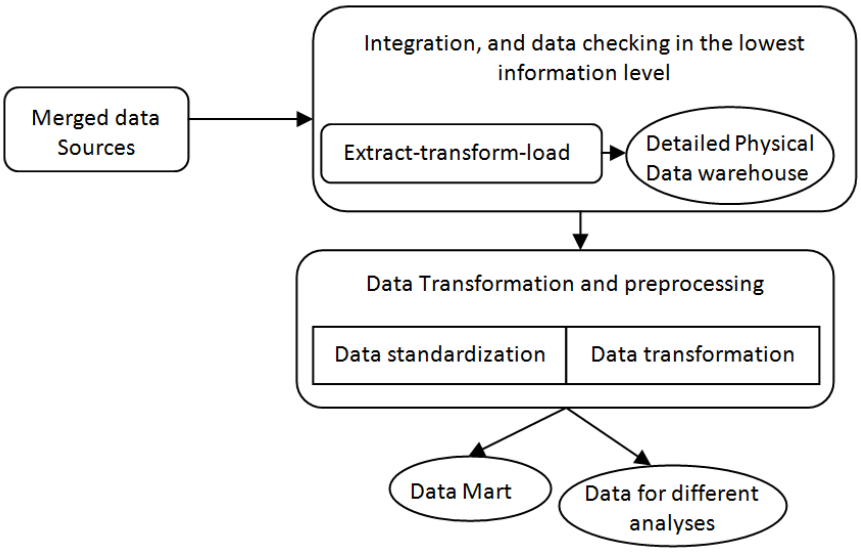

Figure 5: The information pipeline of data integration and preprocessing

Capabilities of the proposed system will be

- The ability to scale to large volumes of data
- Consistent, fast query response times that will allow for iterative speed-of-thought analysis
- Integrated metadata that links the OLAP server and the data warehouse relational database
- The ability to automatically drill from summary and calculated data, which is managed by the OLAP server, in the data warehouse relational database
- A calculation engine that includes robust mathematical functions for computing derived data (aggregations, matrix calculations, cross-dimensional calculations, OLAP-aware formulas and procedural calculations)
- Integration of historical and derived data
- A multi-user read/write environment to support what-if analysis, modeling and planning
- Robust data-access security and user management
- Availability of a wide variety of viewing and analysis tools to support different user communities

References

Abelló, A., & Martín, C. (2003). The Data Warehouse: An Object-Oriented Temporal Database. *In Proc of XX JORNADAS DE INGENIERÍA DEL SOFTWARE Y BASES DE DATOS, Santander*, 675 – 684.

Abelló, A., Romero, O., Pedersen, T.B., Llavori, R.B., Nebot, V., Cabo, M.J.A. & Simitsis, A. (2015). Using Semantic Web Technologies for Exploratory OLAP: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 27(2), 571-588.

Alaskar, K., & Shaikh, A. (2009). Object Oriented Data Modeling for Data Warehousing (An Extension of UML approach to study Hajj pilgrim's private tour as a Case Study*). Int. Arab J. e-Technol*., 1(2), 37 – 41.

Alexandrov, A., Bergmann, R., Ewen, S., Freytag, J. C., Hueske, F., Heise, A., ... & Naumann, F. (2014). The stratosphere platform for big data analytics. *The VLDB Journal*, 23(6), 939 – 964.

Ali, S., Soh, B. & Torabi, T. (2006). A novel approach toward integration of rules into business processes using an agent-oriented framework. *IEEE Transactions on Industrial Informatics*, 2(3), 145 – 154.

Arpitha, P. (2013). Query Optimization in Data Warehouse. *International Journal of Engineering Research & Technology (IJERT)*, 2(8), 562 – 564.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, *27*(3), 822-844.

Atto, A. M., Trouvé, E., Nicolas, J. M. & Lê, T. T. (2016). Wavelet Operators and Multiplicative Observation Models—Application to SAR Image Time-Series Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 54(11), 6606 – 6624.

Badia, A. & Wagner, A. (2014). Complex SQL predicates as quantifiers. *IEEE Transactions on Knowledge and Data Engineering*, 26(7), 1617 – 1630.

Bajda-Pawlikowski, K., Abadi, D. J., Silberschatz, A. & Paulson, E. (2011). Efficient processing of data warehousing queries in a split execution environment. In Proc. International Conference on Management of data, ACM SIGMOD 1165 – 1176.

Berndt, D. J., Hevner, A. R. & Studnicki, J. (2003). The Catch data warehouse: support for community health care decision-making. *Decision Support Systems*, 35(3), 367 – 384.

Bhise, R.B., Thorat, S.S. & Supekar, A.K. (2013). Importance of Data Mining in Higher Education System. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS), ISSN: 2279-0837, ISBN: 2279-0845.* Volume 6, Issue 6 (Jan. - Feb. 2013), PP 18-21.

Bose, P., Kasabov, N. K., Bruzzone, L. & Hartono, R. N. (2016). Spiking Neural Networks for Crop Yield Estimation Based on Spatiotemporal Analysis of Image Time Series. *IEEE Transactions on Geoscience and Remote Sensing*, 54(11), 6563 – 6573.

Boussadi, A., Caruba, T., Zapletal, E., Sabatier, B., Durieux, P. & Degoulet, P. (2012). A clinical data warehouse-based process for refining medication orders alerts. *Journal of the American Medical Informatics Association*, 19(5), 782 – 785.

Brodsky, A., Shao, G. & Riddick, F. (2016). Process analytics formalism for decision guidance in sustainable manufacturing. *Journal of Intelligent Manufacturing*, 27(3), 561 – 580.

Burstein, F., Silval, D. D., Jelinek, H.F. & Stranieri, A. (2013). Multivariate Data-Driven Decision Guidance for Clinical Scientists. In Data Engineering Workshops (ICDEW), IEEE 29th International Conference, 193-199.

Castro, Y. & Kim, Y. J. (2016). Data mining on road safety: factor assessment on vehicle accidents using classification models. International Journal of Crashworthiness, 21(2), 104 – 111.

Chaiboonsri, P. B. S. K. C., & Chaitip, P. (2009). Forecasting with X-12-ARIMA: International tourist arrivals to India and Thailand. *Applied Studies in Agribusiness and Commerce*, 43-61.

Chaki, N. & Sarkar, B. B. (2010). Virtual data warehouse modeling using petri nets for distributed decision making. Journal of Convergence Information Technology, 5(5), 8 – 21.

Chang, L. (2006). Applying Data Mining to Predict College Admission Yield: A Case Study. NEW DIRECTIONS FOR INSTITUTIONAL RESEARCH, no. 131, Fall 2006 © Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)

Chaovanapoonphol, Y., Lim, C., McAleer, M. & Wiboonpongse. A. (2010). Time Series Modelling of Tourism Demand from the USA, Japan and Malaysia to Thailand. *No. EI 2010-18. Report / Econometric Institute, Erasmus University Rotterdam*. 1–15. Retrieved from http://hdl.handle.net/1765/18601

Chen, L., Zhang, W., Liu, H., Feng, S., Chen, C. P. & Wang, H. (2016). A Space Affine Matching Approach to fMRI Time Series Analysis. *IEEE transactions on Nano bioscience*, 15(5), 468 – 480.

Chen, R. J., Bloomfield, P., & Cubbage, F. W. (2008). Comparing forecasting models in tourism. *Journal of Hospitality & Tourism Research*, *32*(1), 3-21.

Cheng, F. T., Yang, H. C. & Lin, J. Y. (2004). Development of holonic information coordination systems with failure-recovery considerations. *IEEE Transactions on Automation Science and Engineering*, 1(1), 58 – 72.

Cherrington, E. A., Barbier, N., Ploton, P., Vincent, G., Sabatier, D., Berger, U. & Pélissier, R. (2016). Equatorial Forests Display Distinct Trends in Phenological Variation: A Time-Series Analysis of Vegetation Index Data from Three Continents. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(8), 3505 – 3511.

Chhabra, R., Kumar, P. & Pahwa, P. (2016). An approach to Design Object Oriented Data Warehouse. *International Journal of Research and Engineering*, 3(3), 54 – 56.

Chu, Fong-Lin. (2004). Forecasting tourism demand: a cubic polynomial approach. *Tourism Management*, 25, 209 – 218.

Chu, Fong-Lin. (2008). Analyzing and forecasting tourism demand with ARAR algorithm. *Tourism Management*, 29, 1185 – 1196.

Chu, Fong-Lin. (2009). Forecasting tourism demand with ARMA-based methods. *Tourism Management*, 30,740 – 751.

Chujai, P., Kerdprasop, N. & Kerdprasop, K. (2013). Time series analysis of household electric consumption with ARIMA and ARMA models. In Proceedings of the International Multi Conference of Engineers and Computer Scientists, 1, 295 – 300.

Cooley, P. C., Roberts, D., Bakalov, V. D., Bikmal, S., Cantor, S., Costandine, T., ... & Li, S. (2008). The model repository of the models of infectious disease agent study. *IEEE Transactions on Information Technology in Biomedicine*, 12(4), 513 – 522.

Cuggia, M., Garcelon, N., Campillo-Gimenez, B., Bernicot, T., Laurent, J. F., Garin, E., ... & Duvauferrier, R. (2011). Roogle: an information retrieval engine for clinical data warehouse. In MIE. 584 – 588.

Cuzzocrea, A. (2011). Retrieving accurate estimates to OLAP queries over uncertain and imprecise multidimensional data streams. *In International Conference on Scientific and Statistical Database Management*. Springer Berlin Heidelberg, 575 – 576.

Cuzzocrea, A., & Gunopulos, D. (2010). Efficiently computing and querying multidimensional OLAP data cubes over probabilistic relational data. In *East European Conference on Advances in Databases and Information Systems*, 132 – 148.

Danubianu, M., Socaciu, T. & Barila, A. (2009). Some aspects of data warehousing in tourism industry. *The Annals of the "Stefan cel Mare" University of Suceava. Fascicle of the Faculty of Economics and Public Administration*, 9(1 (9)), 290 – 296.

Diallo, O., Rodrigues, J. J., Sene, M. & Lloret, J. (2015). Distributed database management techniques for wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems*, 26(2), 604 – 620.

Dimitriadou, K., Papaemmanouil, O. & Diao, Y. (2016). AIDE: An Active Learning-Based Approach for Interactive Data Exploration. *IEEE Transactions on Knowledge and Data Engineering*, 28(11), 2842 – 2856.

Dindarloo, S. R., & Siami-Irdemoosa, E. (2017). Data mining in mining engineering: results of classification and clustering of shovels failures data. *International Journal of Mining, Reclamation and Environment*, *31*(2), 105-118.

Eder, J., Frank, H., Morzy, T., Wrembel, R. & Zakrzewicz, M. (2000). Designing an Object-Relational Data Warehousing System: Project ORDAWA. *In ADBIS-DASFAA Symposium*, 223 – 227.

Evans, R. S., Lloyd, J. F. & Pierce, L. A. (2012). Clinical use of an enterprise data warehouse. *In AMIA Annual Symposium Proceedings, American Medical Informatics Association*, 189.

Fan, J., Zhang, M., Kok, S., Lu, M. & Ooi, B. C. (2015). CrowdOp: Query optimization for declarative crowdsourcing systems. *IEEE Transactions on Knowledge and Data Engineering*, 27(8), 2078 – 2092.

Fattahi, H., Agram, P. & Simons, M. (2017). A Network-Based Enhanced Spectral Diversity Approach for TOPS Time-Series Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2), 777 – 786.

Fender, P. & Moerkotte, G. (2012). Reassessing top-down join enumeration. *IEEE Transactions on Knowledge and Data Engineering*, 24(10), 1803 – 1818.

Fernandes, P. & Teixeira, J. (2008). Applying the artificial neural network methodology for forecasting the tourism time series. *5th International Scientific Conference BUSINESS and MANAGEMENT*. Vilnius, Lithuania.

Gandhi, S. K. & Jain, S. (2011). Data warehouse design: An object-oriented approach. *Journal of Engineering, Science & Management Education*, 4, 16 – 20.

Goh, C. & Law, R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism Management*, 23, 499 – 510.

Gonzalez, H., Han, J., Cheng, H., Li, X., Klabjan, D. & Wu, T. (2010). Modeling massive RFID data sets: A gateway-based movement graph approach. *IEEE Transactions on Knowledge and Data Engineering*, 22(1), 90 – 104.

Gosain, A. & Mann, S. (2010). Object oriented multidimensional model for a data warehouse with operators. *International Journal of Database Theory and Application*, 3(4), 35 – 40.

Gosain, A. & Mann, S. (2014). Empirical validation of metrics for object oriented multidimensional model for data warehouse. *International Journal of System Assurance Engineering and Management*, 5(3), 262 – 275.

Grabova, O., Darmont, J., Chauchat, J. H., & Zolotaryova, I. (2010). Business intelligence for small and middle-sized entreprises. *ACM SIGMOD Record*, 39(2), 39 – 50.

Hamad, M. M. & Qader, B. A. (2014). Knowledge-driven decision support system based on knowledge warehouse and data mining for market management. *Global Journal of Management And Business Research*, 13(10), 1-9.

Hilaly, H., & El-Shishiny, H. (2008). Recent advances in econometric modeling and forecasting techniques for tourism demand prediction. In *Proceedings of Eurochrie Conference, Dubai*.1-9.

Hu, Z. G., Li, J. P., Hu, L. & Yang, Y. (2015). Research and application of data warehouse and data mining technology in medical field. *In Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 12th International Computer Conference on IEEE,* 457 – 460.

Ifeanyi, N., Oghenevwoke, E., Uchenna, A., Amarachi, N., Chinedum, I. & Nwabueze, O. (2014). The Role of Data Warehousing Concept for Improved Organizations Performance and Decision Making. *International Journal of Computer Science and Mobile*, 3(10), 451 – 455

Ikponmwosa, E. G. & Aliu, I. M. (2014). Time Series Analysis on Students' Patronage of an Electronic Library: Case Study of John Harris Library (E-Learning Centre) University Of Benin. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS)*, 19(11), 05 – 12.

Inmon, W. H. (2005). Building the data warehouse. John Wiley & Sons.

Ismail, Z., Yahaya, S.A. & Efendi, R. (2009). Intervention Model for Analyzing the Impact of Terrorism to Tourism Industry. *Journal of Mathematics and Statistics*, 5(4), 322 – 329.

Jebb, A. T., Tay, L., Wang, W. & Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in psychology*, 6, 1 – 24.

Johnson, T. & Srivatsa, S. K. (2012). A Study on Optimization Techniques and Query Execution Operators. *International Journal of Advanced Research in Computer Science*, 3(3), 228 – 233.

Kamble, S., Desai, A. & Vartak, P. (2015). Data mining and data warehousing for supply chain management. *In Communication, Information & Computing Technology (ICCICT), International Conference on IEEE*, 1 – 6.

Karde, P. & Thakare, V. (2010). Selection of materialized views using query optimization in database management: An efficient methodology. *International Journal of Management Systems*, 2(4), 116 – 130.

Kehlenbeck, M. & Breitner, M. H. (2009). Ontology-based exchange and immediate application of business calculation definitions for online analytical processing. *In International Conference on Data Warehousing and Knowledge Discovery, Springer Berlin Heidelberg*, 298 – 311.

Khachatryan, A., Müller, E., Stier, C. & Böhm, K. (2015). Improving accuracy and robustness of self-tuning histograms by subspace clustering. *IEEE Transactions on Knowledge and Data Engineering*, 27(9), 2377 – 2389.

Kolozali, Ş., Puschmann, D., Bermudez-Edo, M. & Barnaghi, P. (2016). On the Effect of Adaptive and Nonadaptive Analysis of Time-Series Sensory Data. *IEEE Internet of Things Journal*, 3(6), 1084 – 1098.

Krasowski, M. D., Schriever, A., Mathur, G., Blau, J. L., Stauffer, S. L. & Ford, B. A. (2014). Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research. *Journal of pathology informatics*, 6, 45 – 45.

Kumar, M. & Anand, M. (2014). An application of time series ARIMA forecasting model for predicting sugarcane production in India. *Studies in Business and Economics*, 9(1), 81 – 94.

Kumar, S. & Toshniwal, D. (2015). A data mining framework to analyze road accident data. *Journal of Big Data*, 2(1), 1 – 26.

Kumar, S. & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. *Journal of Modern Transportation*, 24(1), 62 – 72.

Li, C., He, B., Yan, N. & Safiullah, M. A. (2014). Set predicates in SQL: Enabling set-level comparisons for dynamically formed groups. *IEEE Transactions on Knowledge and Data Engineering*, 26(2), 438 – 452.

Li, Q. & Zhou, M. (2011). The future-oriented grid-smart grid. *Journal of Computers*, 6(1), 98-105.

Lim, C. & McAleer, M. (2002). Time Series forecasts of international travel demand for Australia. *Tourism Management*, 23, 389 – 396.

Liu, J. Y. C. & Lai, S. Y. (2008). Constructing an Object-relational Data Warehouse Using Semi-automated Methods. *Journal of Information Management*, 15(4), 205 – 225.

Liu, L. M., Bhattacharyya, S., Sclove, S. L., Chen, R. & Lattyak, W. J. (2001). Data mining on time series: an illustration using fast-food restaurant franchise data. *Computational Statistics & Data Analysis*, 37(4), 455 – 476.

Loyola, R. C., Sepulveda, A. U. & Hernandez, M. W. (2015). Optimization slowly changing dimensions of a data warehouse using object-relational. *In Chilean Computer Science Society (SCCC), 2015 34th International Conference of the IEEE*, 1 – 6.

Mirge, V., Verma, K. & Gupta, S. (2016). Dense traffic flow patterns mining in bi-directional road networks using density based trajectory clustering. *Advances in Data Analysis and Classification*, 1 – 15.

Mul, D., M., Alons, P., Van der Velde, P., Konings, I., Bakker, J. & Hazelzet, J. (2012). Development of a clinical data warehouse from an intensive care clinical information system. *Computer methods and programs in biomedicine*, 105(1), 22 – 30.

Myalapalli, V. K. & Dussa, K. (2015). Optimizing SQL queries in OLAP database systems. In Information Processing (ICIP), *International Conference on IEEE*, 833 – 838.

Namrata & Singh, S.B. (2014). A Comprehensive Study on Data Warehouse, OLAP and OLTP Technology. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 2(11), 185-190.

Naudé, W. A., & Saayman, A. (2005). Determinants of tourist arrivals in Africa: a panel data regression analysis. *Tourism Economics*, *11*(3), 365-391.

Nguyen, T. M., Schiefer, J. & Tjoa, A. M. (2005). Sense & response service architecture (SARESA): an approach towards a real-time business intelligence solution and its use for a fraud detection

application. *In Proceedings of the 8th ACM international workshop on Data warehousing and OLAP*, ACM, 77 – 86.

Ordonez, C. & Chen, Z. (2009). Evaluating Statistical Tests on OLAP Cubes to Compare Degree of Disease. *IEEE Transactions on Information Technology in Biomedicine*. 13(5), 756 – 765.

Pahwa, P. & Chhabra, R. (2014). An Object Oriented Data Warehouse Design. *International Journal of Soft Computing and Engineering (IJSCE)*, IV (2), 2231 – 2307.

Palaniappan, S. & Lin, C.S. (2008). Clinical Decision Support Using OLAP With Data Mining. *IJCSNS International Journal of Computer Science and Network Security*, 8(9), 290 – 296.

Panov, P., Soldatova, L. & Džeroski, S. (2014). Ontology of core data mining entities. *Data Mining and Knowledge Discovery*, 28(5 - 6), 1222 – 1265.

Papatheodorou, A., & Song, H. (2005). International tourism forecasts: time-series analysis of world and regional data. *Tourism Economics*, *11*(1), 11-23.

Park, D., Yu, J., Park, J. S. & Kim, M. S. (2013). NetCube: a comprehensive network traffic analysis model based on multidimensional OLAP data cube. *International Journal of Network Management*, 23(2), 101 – 118.

Patel, D. & Patel, P. (2015). A Review Paper on Different Approaches for Query Optimization using Schema Object base View. *International Journal of Computer Applications*, 114(4), 16 – 18.

Patelis, A., Petropoulos, C., Nikolopoulos, K., Lin, B., & Assimakopoulos, V. (2005). Tourism planning decision support within an e-government framework. *Electronic Government, an International Journal*, *2*(2), 134-143.

Pedersen, T. B. & Jensen, C.S. (1998). Research issues in clinical data warehousing. *Proceedings Tenth International Conference on Scientific and Statistical Database Management* (Cat. No.98TB100243), Capri, 43 – 52.

Pedersen, T. B. & Jensen, C.S. (1999). Multidimensional data modeling for complex data. *Proceedings 15th International Conference on Data Engineering* (Cat. No. 99CB36337), Sydney, NSW, 336-345. doi: 10.1109/ICDE.1999.754949

Peng, Y., Zhang, Y., Tang, Y. & Li, S. (2011). An incident information management framework based on data integration, data mining, and multi-criteria decision making. *Decision Support System*, 51(2), 316 – 327.

Peng, Z., Li, Q., Feng, L., Li, X. & Liu, J. (2005). Using object deputy model to prepare data for data warehousing. *IEEE Transactions on Knowledge and Data Engineering*, 17(9), 1274 – 1288.

Qiao, M., Cheng, H., Chang, L. & Yu, J. X. (2014). Approximate shortest distance computing: A query-dependent local landmark scheme. *IEEE Transactions on Knowledge and Data Engineering*, 26(1), 55 – 68.

Rahul, S. & Janardan, R. (2014). A General Technique for Top-$ k $ Geometric Intersection Query Problems. *IEEE Transactions on Knowledge and Data Engineering*, 26(12), 2859 – 2871.

Ram, S. & Doegar, A. (2015). A Comparative Study of Data Mining Techniques for Predicting Disease Using Statlog Heart Disease Database. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(6), 1 – 5.

Ramachandra, K., Chavan, M., Guravannavar, R. & Sudarshan, S. (2015). Program transformations for asynchronous and batched query submission. *IEEE Transactions on Knowledge and Data Engineering*, 27(2), 531 – 544.

Ramamurthy, K. R., Sen, A. & Sinha, A. P. (2008). An empirical investigation of the key determinants of data warehouse adoption. *Decision support systems*, 44(4), 817 – 841.

Ravat, F. & Teste, O. (2000). A temporal object-oriented data warehouse model. *In International Conference on Database and Expert Systems Applications, Springer Berlin Heidelberg*, 583 – 592.

Roelofs, E., Persoon, L., Nijsten, S., Wiessler, W., Dekker, A. & Lambin, P. (2013). Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. *Radiotherapy and Oncology*, 108(1), 174 – 179.

Ross, T. R., Ng, D., Brown, J. S., Pardee, R., Hornbrook, M. C., Hart, G. & Steiner, J. F. (2014). The HMO Research Network Virtual Data Warehouse: a public data model to support collaboration. EGEMS 2 (1): 1049-1056.

Sahama, T. R. & Peter, C. (2007). A Data Warehouse Architecture for Clinical Data Warehousing. In Roddick, J. F. and Warren, J. R., Eds. *Proceedings Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007) CRPIT*, 68, Ballarat, Victoria, 227 – 232.

Sahay, B. S. & Ranjan, J. (2008). Real time business intelligence in supply chain analytics. *Information Management & Computer Security*, 16(1), 28 – 48.

Salguero, A., Araque, F., Delgado, C. (2008). Spatio-temporal ontology based model for Data Warehousing. *7th WSEAS Int. Conf. on TELECOMMUNICATIONS and INFORMATICS (TELE-INFO '08)*, Istanbul, Turkey. 125-130.

Sarkar, A., Choudhury, S., Chaki, N. & Bhattacharya, S. (2009). Conceptual Level Design of Object Oriented Data Warehouse: Graph Semantic Based Model. *INFOCOMP Journal of Computer Science*, 8(4), 60 – 70.

Sharma, L. & Mehta, N. (2012). Data mining techniques: A tool for knowledge management system in agriculture. *International journal of scientific & technology research*, 1(5), 67 – 73.

Shitan, M. & Wee, P.M.J. (2003a). ARAR and Long Memory Modelling of Tourist Arrivals to Malaysia. *Proceedings of National Conference on Management Science and Operations Research*, 1, 258 – 271.

Shitan, M. & Wee, P.M.J. (2003b). A comparison between ARFIMA Model and Regression Model with ARMA Errors for Tourist Arrivals in Malaysia, *Proceedings of National Seminar on Decision Science*, Langkawi, Kedah, 109 – 114.

Shnitzer, T., Talmon, R. & Slotine, J. J. (2017). Manifold Learning With Contracting Observers for Data-Driven Time-Series Analysis. *IEEE Transactions on Signal Processing*, 65(4), 904 – 918.

Sim, H., Choi, D. & Kim, C. O. (2014). A data mining approach to the causal analysis of product faults in multi-stage PCB manufacturing. *International Journal of Precision Engineering and Manufacturing*, 15(8), 1563 – 1573.

Singh, R. & Singh, K. (2010). A descriptive classification of causes of data quality problems in data warehousing. *International Journal of Computer Science Issues*, 7(3), 41 – 50.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. *Tourism management*, *29*(2), 203-220.

Soni, J., Ansari, U., Sharma, D. & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8), 43 – 48.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, *2*(02), 250-255.

Stell, A., Sinnott, R. & Ajayi, O. (2006). Secure Federated Data Retrieval in Clinical Trials. *Proc. IASTED International Conference on Telehealth*, Banff, Alberta, Canada. 1 – 8.

Stolba, N., & Tjoa, A. M. (2006). The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making. *International Journal of Computer Systems Science and Engineering*, *3*(3), 143-148.

Stonebraker, M., & Brown, P. (1999). Object-Relational DBMSs: tracking the next great wave. San. Francisco: Mogan Kaufmann Pub.

Sukheja, D. & Singh, U. K. (2010). Query optimizer model for performance enhancement of data mining based query. *International journal of computer science & communication,* 1(1), 235 – 237.

Tang, B., He, H., Baggenstoss, P. M. & Kay, S. (2016). A Bayesian classification approach using class-specific features for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 28(6), 1602 – 1606.

Taylor, P., Griffiths, N., Bhalerao, A., Anand, S., Popham, T., Xu, Z. & Gelencser, A. (2016). Data mining for vehicle telemetry. *Applied Artificial Intelligence*, 30(3), 233 – 256.

Van, G. A., Staals, F., Löffler, M., Dykes, J. & Speckmann, B. (2017). Multi-Granular Trend Detection for Time-Series Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1), 661 – 670.

Viswanathan, S., Widiarta, H. & Piplani, R. (2008). Forecasting aggregate time series with intermittent subaggregate components: top-down versus bottom-up forecasting. *IMA Journal of Management Mathematics*, 19, 275 – 287.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P. & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 1 – 35.

Watson, H. J. (2002). Recent developments in data warehousing. *Communications of the Association for Information Systems*, *8*(1), 2289 – 2292.

Witt, S F. (2009). Time Varying Parameter Structural Time Series Models: An Application to Tourism Demand. *The 29th International Symposium on Forecasting*, Hongkong. Retrieve from www.forecasters.org/isf.

Xu, Z., Tu, Y. C. & Wang, X. (2015). Online Energy Estimation of Relational Operations in Database Systems. *IEEE Transactions on Computers*, 64(11), 3223 – 3236.

Yang, Y., Zou, H. Combing Time Series Models for Forecasting. *Journal of American Statistical Association*, 96, 574 – 588.

Ye, Z., Mistry, S., Bouguettaya, A. & Dong, H. (2016). Long-term QoS-aware cloud service composition using multivariate time series analysis. *IEEE Transactions on Services Computing*, 9(3), 382 – 393.

Yokoya, N., Zhu, X. X. & Plaza, A. (2017). Multisensor coupled spectral unmixing for time-series analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5), 2842 – 2857.

Yorucu, V. (2003). The Analysis of Forecasting Performance by Using Time Series Data for Two Mediterranean Islands. *Review of Social, Economic & Business Studies, FBE Journal*, 2, 175 – 196.

Zhang, X., Zhang, X., Liu, H. & Liu, X. (2016). Multi-Task Multi-View Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 28(12), 3324 – 3338.

Zhou, X., Chen, S., Liu, B., Zhang, R., Wang, Y., Li, P., ... & Yan, X. (2010). Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. *Artificial Intelligence in Medicine*, 48(2), 139 – 152.

Zhou, X., Liu, B., Wang, Y., Zhang, R., Li, P., Chen, S., ... & Zhang, H. (2008). Building clinical data warehouse for traditional Chinese medicine knowledge discovery. In *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, 1, 615 – 620. IEEE.